21 RESOURCES TO FIND ALL THE DATA YOU NEED

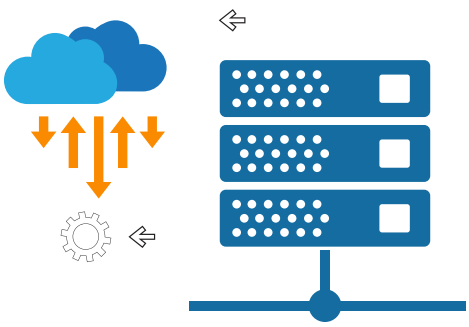
simplilearn

TABLE OF CONTENTS

WHAT IS THE MOST IMPORTANT THING YOU WILL NEED?

But where are you going to find it?

Finding the right data set can be the most difficult step of producing a visualization or building an application. Data visualization may be the final product, but it involves various intermediate steps, which include finding reliable data, getting this data in to the right format, cleaning the data, and then finding a way to effectively develop the story that you had initially visualized.



Finding a good source for all your data solves half the problem. Here's a list of free sources to find all the data you're going to need, no matter your project!

1 UNIVERSITY LIBRARIES

What is the first thing you think of when you want information? The good old library, of course!

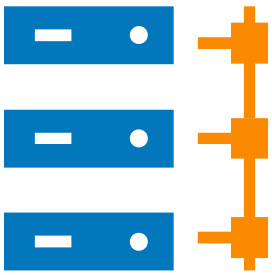
Today, most libraries maintain electronic records and online databases, making it easy to find exactly what you're looking for. University statistics departments also tend to keep an index of the data somewhere.

Here are the most popular university libraries commonly used by data scientists and researchers -

DATA AND STORY LIBRARY

Full of stories and datafiles which illustrate the use of basic statistics methods, this awesome site is a godsend for anyone looking to find the right stat to make your point, from students working on their projects to statisticians and analytics professionals.

The library is more popularly called DASL, pronounced ‘dazzle’.



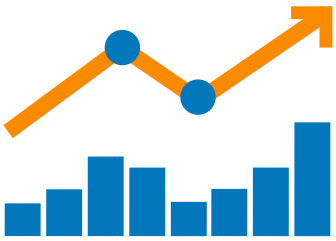
BERKELEY DATA LAB

The Berkeley Data Lab offers consultations on research that involves numeric data, which includes finding and recommending data sources and advising on technical issues such as file format conversion, web scraping, and basic statistical software.

The Lab also provides workstations with analytical software such as ArcGIS, Stata, SAS, SPSS, T, and Python.

UCLA STATISTICS DATA SETS

UCLA Statistics Data Sets is a collection of datasets that are exclusively from UCLA researchers and are utilized for a variety of classroom applications.



2 DATA AGGREGATORS

If you are unable to find a specific source that meets your data needs, aggregators could make your job a lot easier. Data aggregator websites index lists of data sources, categorized by specialization, purpose, and more.

Here are a few data aggregators that are popular among professionals from all backgrounds -

FREEBASE

Freebase is an online collection of structured data that is harvested from various sources, including individual, user-submitted wiki contributions.

NUMBRARY

Numbrary's data is all open-source and is based on information from various US governmental departments.

INFOCHIMPS

Infochimps allows easier ways to build and manage Big Data systems behind the applications to quickly deliver actionable insights. With their cloud solutions, users will benefit more from the fastest way to deploy Big Data applications in complex, hybrid cloud environments

AMAZON PUBLIC DATA SETS

The public data sets on AWS provide a centralized repositories of data which can be integrated seamlessly into AWS cloud based applications. If you are an AWS user or are looking to use AWS's cloud solutions at some point in the future, this data source is indispensable.

DATA MARKET

This is a good place if you are looking to explore data in healthcare, economics, agriculture, and the automotive industry.

3 GOVERNMENT WEBSITES

With the modern approach to administration, there is fresh emphasis laid on free availability of data and transparency. This has meant that a lot of government organizations are providing data in the public domain. With the launch of of data.gov, much of all of this data is now available in one place. There is also a huge number of non-governmental sites that are aiming to make politicians more accountable by uploading performance, fund utilization, and other information important to a tax-payer.

Here are the most popular governmental data sources:

CENSUS BUREAU

The Census bureau houses a ton of information about our lives around income, race, education, population and business.

DATA.GOV

Data.gov is THE go-to government resource if you are looking for data pertaining to the government. The site claims to have up to 400,000 data sets, which are both raw and geo spatial, in a variety of formats.

SOCRATA

Socrata is another great place to explore archives of government data. Their inbuilt data visualization tools make the process of sifting through voluminous piles of data smoother and more interesting.

4 NEWS PORTALS, DIGITAL CONTENT PUBLICATIONS

Ever come across an infographic published by a media house of a news outlet? If you haven't noticed yet, major news organizations always include a reference to their sources somewhere on the graphic, or they are at least mentioned in the accompanying article. This source is usually never a direct link, but a quick online search will land you on the page.

There are times when you may have to mail someone to get the same data, but these people are usually more than happy that you are showing an interest in their analysis or data.

The most popular data sources you can use are:

CENSUS BUREAU

The New York Times has a fantastic API and a great explorer that can access any publication or article.

THE GUARDIAN DATABLOG

The Guardian datablog regularly and consistently publishes articles with visualizations. Guardian's data sources are also freely accessible with a Google doc.

5 UNITED NATIONS & OTHER INTERNATIONAL ORGANIZATIONS

There are a number of international organizations that keep data about the world – mainly developmental and health indices.

The chief drawback of these sources is that the data is often not standardized across countries.

With this caveat, the most popular sources international data sources you can use are:

>> UN RELATED SITES

The best place to go is the UN and its related sites. The UNICEF and the World Health Organization are extremely rich in all sorts of data, from world hunger statistics to mortality rates.

6 SPORTS PORTALS

Who doesn't love sport? The average American spends around 5 hours a day discussing sports and leisure activities related to sports, with an insatiable appetite for blow-by-blow sports news updates.

Sports portals thus see huge volumes of traffic and data generated every day.

But where can you access this data?

Apart from mags like Sports Illustrated, here are a few websites where you can expect to find compiled data for the specific sport they relate to.

BASKETBALL REFERENCE | BASEBALL DATABANK | DATABASE FOOTBALL

ESPN has also recently come up with their own API.

7 GEOGRAPHICAL DATA

Say you have a mapping software. Where are you going to find the geographical data that you need?

Worry not!

You are in luck. There is a vast amount of shapefiles at your disposal.

The most popular sources that you can use are:

TIGER

TIGER is an acronym for Topologically Integrated Geographic Encoding and Referencing. Used by the United States Census Bureau, this format describes land attributes (buildings, rivers, roads, and lakes) as well as areas such as census tracts. TIGER was developed to improve the Bureau's Decennial Census process.

WEATHERBASE

Weatherbase gives detailed statistics on temperature, humidity, and precipitation for close to 27,000 cities.

OPENSTREETMAP

OpenStreetMap (OSM) is a collaborative project aimed at creating a free, editable world map, and is considered a relevant example of volunteered geographic information.

WUNDERGROUND

Wunderground has detailed sets of information on the weather and also lets you look up historical data by city or zip code. It provides data on the wind, temperature, precipitation, and hourly observations for the day

8 GET IT FROM AN API

Many sites and applications make their data freely available through APIs. Twitter, Google, and Yahoo are a few popular names that have their own APIs.

For a more detailed catalog, visit Programmable Web.

9 SCRAPE THE DATA

If all else fails, scrape your data. It is easy to find a site that serves data through the HTML pages, and then scrape the data with JavaScript, Python, or any other language you are comfortable with.

Take your data skills to the next level with a certification in Data Science. Learn about data visualization, exploration and statistical concepts like linear and logistic regression, cluster analysis, forecasting, and programming in R and boost your skills.

With 30+ hours of training in R Language, real life case studies, industry project experience in diverse fields, and simulation exams, you will be all set to take on the intimidating world of data. Make a name for yourself as a data scientist, a profession Harvard Business Review calls the sexiest of the 21st century.

Features of Simplilearn's Data Science with R Language Online Training include:

30+ hours of training in R language

2 Data Science simulation exams

8 real life case studies

20 hours of industry project experience in diverse fields

Hands-on experience in predictive modeling on R